$Y = X\beta + \varepsilon$, $\varepsilon \sim N(0, I_n)$

n $n \times p$ p n

$\hat{\beta}^{(OLS)} = \underset{\beta}{\text{argmin}} \, \|Y - X^T\beta\|_2^2$

$\qquad = (X^TX)^{-1}X^TY$

① Risk of estimators

② Distribution of $\hat{\beta}^{(PCR)}$

Principal Component Regression

$\underline{PCA}$: $X = \begin{bmatrix} X_1^T \\ \vdots \\ X_n^T \end{bmatrix} \in \mathbb{R}^{n \times p}$, $\frac{1}{n}\sum_{i=1}^{n} X_i = 0$, $S = \frac{1}{n} X^TX = PDP^T$

P orthogonal, $P^TP = I_p$, $P = [v_1 \cdots v_p]$, $D = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_p \end{bmatrix}$, $\begin{array}{l}\lambda_1 \geq \ldots \geq \lambda_p \\ \lambda_i \geq 0\end{array}$

Take k largest λ_i, $P_k = [v_1, \cdots, v_k]$, $Z = XP_k$, $Z_i = P_k^T X_i$

$\underline{PCR}$: $\hat{\gamma} = \underset{\gamma \in \mathbb{R}^k}{\text{argmin}} \, \|Y - XP_k\gamma\|_2^2 = (P_k^TX^TXP_k)^{-1}P_k^TX^TY$

$\hat{\beta}^{(PCR)} = P_k\hat{\gamma} = P_k(P_k^TX^TXP_k)^{-1}P_k^TX^TY$

scree plot

$$Y = X\beta + \varepsilon, \quad \varepsilon \sim N(0, I_n)$$
$$\underset{n}{} \quad \underset{n\times p}{} \quad \underset{p}{} \quad \underset{n}{}$$

$$\hat{\beta}^{(OLS)} = \underset{\beta}{\arg\min} \|Y - X^T\beta\|_2^2$$
$$= (X^TX)^{-1}X^TY$$

$$\hat{\beta}^{(PCR)} = P_k(P_k^TX^TXP_k)^{-1}P_k^TX^TY$$

① Risk of estimators

② Distribution of $\hat{\beta}^{(PCR)}$

bias $\qquad$ variance

OLS

PCR

① $\hat{\beta}$ estimator for β, $\mathbb{E}[\hat{\beta}] = \mu$, $Cov(\hat{\beta}) = \Sigma$
$\hat{\beta}(Y)$

$$\mathbb{E}[\|\hat{\beta} - \beta\|_2^2] = \mathbb{E}[\|\hat{\beta} - \mu + \mu - \beta\|_2^2]$$
$$= \mathbb{E}[(\hat{\beta} - \mu + \mu - \beta)^T(\hat{\beta} - \mu + \mu - \beta)]$$
$$= \underbrace{\mathbb{E}[\|\hat{\beta} - \mu\|_2^2]}_{\text{variance}} + 2\underbrace{\mathbb{E}[(\hat{\beta} - \mu)^T(\mu - \beta)]}$$

$$= \mathbb{E}[\hat{\beta} - \mathbb{E}[\hat{\beta}]]^T(\mu - \beta)$$
$$= (\mathbb{E}[\hat{\beta}] - \mathbb{E}[\hat{\beta}])^T(\mu - \beta)$$
$$= 0$$

$$+ \underbrace{\|\beta - \mu\|_2^2}_{\|\text{bias}\|_2^2}, \quad \text{bias} = \mu - \beta$$

$$\mathbb{E}[\|\hat{\beta} - \mu\|_2^2] = \mathbb{E}[(\hat{\beta} - \mu)^T(\hat{\beta} - \mu)] = \mathbb{E}[tr((\hat{\beta} - \mu)^T(\hat{\beta} - \mu))]$$
$$= \mathbb{E}[tr((\hat{\beta} - \mu)(\hat{\beta} - \mu)^T)] = tr(\mathbb{E}[(\hat{\beta} - \mu)(\hat{\beta} - \mu)^T]) = tr(\Sigma)$$

$$Y = X\beta + \varepsilon, \quad \varepsilon \sim N(0, I_n)$$

$n \times p \quad p \quad n \quad X \text{ full rank}$

$$\hat{\beta}^{(OLS)} = \underset{\beta}{\arg\min} \|Y - X^T\beta\|_2^2$$

$$= (X^T X)^{-1} X^T Y$$

$$\hat{\beta}^{(PCR)} = P_k(P_k^T X^T X P_k)^{-1} P_k^T X^T Y$$

① Risk of estimators

② Distribution of $\hat{\beta}^{(PCR)}$

$$\hat{\beta}^{(PCR)} \sim N(P_k P_k^T \beta, \; P_k D_k^{-1} P_k^T)$$

$\qquad\qquad\qquad$ bias $\qquad\qquad$ variance

OLS

PCR

---

② $\hat{\gamma}_k = \|Y - X P_k \gamma\|_2^2$, $\quad \beta^{(PCR)} = P_k \hat{\gamma}$

$$\hat{\beta}^{(PCR)} = P_k(P_k^T X^T X P_k)^{-1} P_k^T X^T (X\beta + \varepsilon)$$

$$= P_k(P_k^T P D P^T P_k)^{-1} P_k^T (X^T X\beta + X^T \varepsilon)$$

$$= \begin{bmatrix} v_1^T \\ \vdots \\ v_k^T \end{bmatrix} \begin{bmatrix} v_1 \cdots v_n \end{bmatrix} = \begin{bmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{bmatrix} \underbrace{\quad}_{k} \; 0 \quad \underbrace{\qquad}_{p-k}$$

$$\left[P_k^T P D P^T P_k = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_k \end{bmatrix} = D_k \right]$$

$$= P_k D_k^{-1} P_k^T (X^T X\beta + X^T \varepsilon)$$

$$\mathbb{E}[\hat{\beta}^{(PCR)}] = P_k D_k^{-1} \underbrace{P_k^T P D P^T}_{} \beta = P_k \begin{bmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{bmatrix} 0 \bigg] P^T \beta = P_k P_k^T \beta$$

$$= \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_k \end{bmatrix} 0 \bigg]$$

$$\mathrm{Cov}(\hat{\beta}^{(PCR)}) = P_k D_k^{-1} P_k^T X^T X P_k D_k^{-1} P_k^T$$

$$= P_k D_k^{-1} \underbrace{P_k^T P D P^T P_k}_{= D_k} D_k^{-1} P_k^T = P_k D_k^{-1} P_k^T$$

$Y = X\beta + \varepsilon$, $\varepsilon \sim N(0, I_n)$

n $n \times p$ p n X full rank

$\hat{\beta}^{(OLS)} = \underset{\beta}{\arg\min} \|Y - X^T\beta\|_2^2$

$\qquad = (X^TX)^{-1}X^TY$

$\hat{\beta}^{(PCR)} = P_k(P_k^TX^TXP_k)^{-1}P_k^TX^TY$

① Risk of estimators

② Distribution of $\hat{\beta}^{(PCR)}$

$\hat{\beta}^{(PCR)} \sim N(P_kP_k^T\beta, P_kD_k^{-1}P_k^T)$

③

	bias	variance
OLS	0	$\sum_{i=1}^{p} \frac{1}{\lambda_i}$
PCR	$P_kP_k^T\beta - \beta$ $= (I - P_kP_k^T)\beta$	$\sum_{i=1}^{k} \frac{1}{\lambda_i}$

Ridge

③ $\hat{\beta}^{(OLS)} \sim N(\beta, \underbrace{(X^TX)^{-1}}_{= PDP^T})$

$\mathbb{E}[\|\hat{\beta}^{(OLS)} - \beta\|_2^2] = tr(PD^{-1}P^T) = tr(\underbrace{P^TP}_{I_p}D^{-1}) = tr(D^{-1})$

$\qquad\qquad = \sum_{i=1}^{p} \frac{1}{\lambda_i}$

$\mathbb{E}[\|\hat{\beta}^{(PCR)} - \mathbb{E}[\hat{\beta}^{(PCR)}]\|_2^2] = tr(P_kD_k^{-1}P_k^T)$

$\qquad = tr(P_k^TP_kD_k^{-1}) = tr(D_k^{-1}) = \sum_{i=1}^{k} \frac{1}{\lambda_i}$

$[\quad]\begin{bmatrix} \\ \end{bmatrix} = I_k$

$$Y = X\beta + \varepsilon, \quad \varepsilon \sim N(0, I_n)$$

n ny p n X full rank

$$\hat{\beta}^{(OLS)} = \underset{\beta}{\text{argmin}} \|Y - X^T\beta\|_2^2$$

$$= (X^TX)^{-1}X^TY$$

$$\hat{\beta}^{(PCR)} = P_k(P_k^TX^TXP_k)^{-1}P_k^TX^TY$$

① Risk of estimators

② Distribution of $\hat{\beta}^{(PCR)}$

$$\hat{\beta}^{(PCR)} \sim N(P_kP_k^T\beta, P_kD_k^{-1}P_k^T)$$

③

	bias	variance
OLS	0	$\sum_{i=1}^{p} \frac{1}{z_i}$
PCR	$P_kP_k^T\beta - \beta$ $= (I - P_kP_k^T)\beta$	$\sum_{i=1}^{k} \frac{1}{z_i}$
Ridge	$P\tilde{D}P^T\beta - \beta$	$\sum_{i=1}^{p} \frac{z_i}{(z_i + \tau)^2}$

③ ctd. $\hat{\beta}^{(R)} = \underset{\beta \in \mathbb{R}^p}{\text{argmin}} \|Y - X\beta\|_2^2 + \tau\|\beta\|_2^2$

$$= (X^TX + \tau I_p)^{-1}X^TY$$

$$E[\hat{\beta}^{(R)}] = (X^TX + \tau I_p)^{-1}X^TX\beta = (PDP^T + \tau I_p)^{-1}PDP^T\beta$$

$$= (P(D + \tau I_p)P^T)^{-1}PDP^T\beta = P(D + \tau I_p)^{-1}P^TPDP^T\beta$$

$$= P\tilde{D}P^T\beta, \quad \tilde{D}_{ii} = \frac{z_i}{z_i + \tau} = \frac{1}{1 + \frac{\tau}{z_i}} \quad \text{"Shrinkage"}$$

$$Var(\hat{\beta}^{(R)}) = tr\left[(X^TX + \tau I_p)^{-1}X^TX(X^TX + \tau I_p)^{-1}\right]$$

$$= tr\left(P(D + \tau I)^{-1}P^TPDP^TP(D + \tau I)^{-1}P^T\right)$$

$$= tr(P\tilde{\tilde{D}}P^T) = tr(\tilde{\tilde{D}}), \quad \tilde{\tilde{D}}_{ii} = \frac{z_i}{(z_i + \tau)^2}$$

$$= \sum_{i=1}^{p} \frac{z_i}{(z_i + \tau)^2}$$